International Journal of Human-Computer Studies 124 (2019) 56–66

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Webthetics: Quantifying webpage aesthetics with deep learning

Qi Dou[b], Xianjun Sam Zheng[*,a], Tongfang Sun[c], Pheng-Ann Heng[b]

[a] Faculty of Psychology, Beijing Normal University, Beijing, China
[b] Department of Computer Science and Engineering, The Chinese University of Hong Kong, HK
[c] Human Centered Design and Engineering, University of Washington, Seattle, WA, USA

ARTICLE INFO

Keywords:
Webpage aesthetics
Deep learning
Web visual design
User experience

ABSTRACT

As web has become the most popular media to attract users and customers worldwide, webpage aesthetics plays an increasingly important role for engaging users online and impacting their user experience. We present a novel method using deep learning to automatically compute and quantify webpage aesthetics. Our deep neural network, named as *Webthetics*, which is trained from the collected user rating data, can extract representative features from raw webpages and quantify their aesthetics. To improve the model performance, we propose to transfer the knowledge from image style recognition task into our network. We have validated that our method significantly outperforms previous method using hand-crafted features such as colorfulness and complexity. These promising results indicate that our method can serve as an effective and efficient means for providing objective aesthetics evaluation during the design process.

1. Introduction

With the rapid development of Internet since its inception decades ago, web has become the most popular media for information searching, company marketing, entertainment and social activities (Hoffmann and Krauss, 2004). Previous studies have shown that users make a rapid but lasting impression on attractiveness of webpage within as short as 50 ms (Lindgaard et al., 2006; Tractinsky et al., 2006). This extraordinarily rapid impression is made, to some extent, based on visual aesthetics and can highly impact user experience online (Reinecke et al., 2013; Zheng et al., 2009). For example, positive impressions can improve users' trust level of the website and also encourage positive user behaviors such as longer engagement and eventually increasing level of purchase intention (Lu et al., 2013). Previous studies have also experimentally demonstrated that the visually appealing is important for first-impression judgments and requires consideration during webpage design (Lindgaard and Dudek, 2003; Lindgaard et al., 2011). Hence, webpage aesthetics has become an increasingly key factor in web design considerations (Schmidt et al., 2009).

Professional design is quite a time and labor intensive process where designers build up various elements based on heuristic guidelines or established principles (Galitz, 2007; Zheng et al., 2009). In current practice, the assessment of webpage aesthetics quality is depended on the designers' experience and intuitive judgments. However, this way of evaluation is subjective and may be influenced by the designer's background, style and taste. Moreover, existing studies have also shown that professional designers may not always share the same impressions with their targeted users (Heer and Bostock, 2010; Park et al., 2004). Another approach is to conduct questionnaires or surveys in order to collect users' aesthetic feedback of the design (Hassenzahl, 2004). This method can obtain wider and more objective evaluations of design aesthetics, but it is often expensive and time consuming to perform the survey, collect and analyze the data. Therefore, it is necessary to develop an automatic and objective method for aesthetics evaluation, which would help to improve webpage design aesthetics.

Meanwhile, the fact that users make reliable aesthetic judgments within a fraction of second suggests that there exists a highly optimized aesthetics assessment mechanism within the biological visual system (Hubel and Wiesel, 1979; Treisman and Gelade, 1980). The human brain should be able to generate representations that can describe visual attractiveness to a large extent, although the judgment of aesthetics also has subjective, sophisticated and personalized aspects (Reinecke et al., 2013; Wu et al., 2011).

Previous researches have attempted to use hand-crafted features to quantify webpage aesthetics (Michailidou et al., 2008; Reinecke et al., 2013; Wu et al., 2011; Zheng et al., 2009). Those features are heuristically defined to reflect important aspects of web visual design guidelines. Usually, the webpage design is decomposed into many interrelated factors such as colorfulness, balance, complexity, symmetry,

* Corresponding author.
E-mail address: sam.zheng@gmail.com (X.S. Zheng).

https://doi.org/10.1016/j.ijhcs.2018.11.006
Received 3 June 2017; Received in revised form 4 October 2018; Accepted 9 November 2018
Available online 27 November 2018
1071-5819/

etc. However, how these factors work together to contribute to aesthetics perception can be quite complex, and previous studies often examined these factors independently to avoid modeling the complexity. However, considering various design factors holistically as an integration is exactly what professional designers do in everyday practice (Guindon, 1990; Jaarsveld and Leeuwen, 2005; Lawson, 2006). It is inadequate to perform aesthetics predictions only with hand-crafted features and linear or simple non-linear models like what previous methods did.

Deep learning techniques have revolutionized many research and industry areas in the past several years (Hinton and Salakhutdinov, 2006; LeCun et al., 2015), and the convolutional neural networks (CNN) in particular have continuously achieved superior performance and have become the state-of-the-arts method in computer vision (He et al., 2016; Krizhevsky et al., 2012). The neural networks are capable of automatically extracting high-level features directly from raw input data via a hierarchical architecture composing high non-linearities. In this paper, we propose a novel method to quantify the webpage aesthetics with deep convolutional neural networks, by bridging the gap between human computer interaction research and web design practice and artificial intelligence methodologies. We refer to our proposed model as *Webthetics*. Our contributions are as follows:

- We develop a novel deep learning based method to automatically quantify the aesthetics of web visual design. We explore knowledge transfer from image style recognition task into our aesthetics evaluation task, which significantly improves the performance of our model. Experimental results show that our predicted aesthetics ratings are highly correlated with collected user rating data.
- We further validate our proposed method by comparing with previous method using hand-crafted features of colorfulness and complexity, and our results have outperformed the baseline by a significant margin. The promising results demonstrate the effectiveness of deep learning for webpage design aesthetics quantification, and potentially for human computer interaction.
- We also conduct empirical experiments illustrating that our deep learning model is sensitive to some manipulation factors including layout, balance, content information and spatial frequency. These findings imply that our method has potential to serve as an efficient and effective tool for providing objective aesthetics evaluation during web design process.

In the following, we first provide a literature review. Then, we introduce our used dataset and employed evaluation measurement in this paper. Afterwards, we explain in details of our aesthetics quantification method based on deep learning and present our experimental results. Finally, we close with a discussion, a conclusion and a deliberation of future work. To facilitate future researches, the implementation of our method is publicly available at: https://github.com/carrenD/Webthetics

2. Related work

2.1. Computational website aesthetics

Great efforts have been made in the CHI community to study the importance of webpage aesthetics (Thorlacius, 2007; Tuch et al., 2012), to explore the webpage visual design (Harrison et al., 2015; Silvennoinen and Jokinen, 2016; Zhang and Kong, 2010), and to evaluate the aesthetics of web design (Michailidou et al., 2008; Reinecke et al., 2013; Wu et al., 2011; Zheng et al., 2009). For example, the Michailidou et al. (2008) presented an investigation into user perception of the visual complexity and aesthetics appearance of webpages. Their results demonstrated a strong and high correlation between users' perception of visual complexity, structural elements and aesthetics of webpages. Computational webpage aesthetics is regarded as a promising direction to provide designers an efficient surrogate for the cost and time intensive user studies during the design process. A prior work from Zheng et al. (2009) studied low-level image statistics to characterize the webpage's organizational symmetry, balance and equilibrium. These computational attributes of webpages were evaluated for the relationship with user participants' ratings on four aesthetic and affective dimensions. Later, the Wu et al. (2011) presented to compute web visual quality based on structural information such as the page layout, text positions and distributions, inner image positions and background areas. They employed the multi-cost-sensitive learning and multi-value regression to assign scores with a model of good generalization capability. A recent work from Reinecke et al. (2013) demonstrated that colorfulness and complexity are two important factors to describe the aesthetics of webpage design. Furthermore, their following work (Reinecke and Gajos, 2014) contributed the first public dataset for webpage visual appealing research, which would greatly support following studies on computational website aesthetics.

This paper shares the same goal with these previous works. We employ neural networks to automatically learn representations from webpages and their corresponding user ratings, rather than using hand-craft features based on experience or from design heuristic considerations. The work (Khani et al., 2016) also proposed to employ deep learning for webpage aesthetics computation. Their method has three steps, i.e., CNNs for feature extraction, principal component analysis for feature dimension reduction, support vector machine for classification. In contrast, our aesthetics computation model is trained end-to-end, i.e., from raw webpage pixels to aesthetics rating predictions. In addition, the Khani et al. formulated the task as a binary classification problem, i.e., classifying the webpage aesthetics into good or bad. We treat the task as a regression problem to explicitly output the rating scores of webpage aesthetics. Further refining the quantification of webpage aesthetics also lies in the future work in Khani et al. (2016).

2.2. Deep learning for photo aesthetics

With the goal of empowering computers with the capability to perceive aesthetics, researchers in computer vision have also made some efforts to automatically estimate aesthetics quality of photographs or natural images. Early works have relied on hand-crafted features which either encapsulated visual design concepts (e.g., colorfulness, saturation, rule of thirds, etc.) (Datta et al., 2006; Dhar et al., 2011; Nishiyama et al., 2011) or utilized generic image descriptors (e.g., SIFT and Fisher Vector) (Marchesotti et al., 2011). With recent advancement of deep learning, researches have established state-of-the-art performance for photograph aesthetics estimation with neural networks (Kang et al., 2014; Kong et al., 2016; Lu et al., 2014). Meanwhile, researches on artistic recognition of photographs have also been explored. Notably, Karayev et al. (2013) have proposed to recognize image styles (e.g., "vintage" and "romantic") with the deep CNNs and presented sound performance.

These works have successfully demonstrated the outstanding effectiveness of deep learning for photo aesthetics quantification. These results have confirmed the feasibility to automatically learn certain features which can universally describe aesthetics and attract majority users, although the compositional difference between webpages and photographs cannot be neglected and need to be studied carefully (Reinecke et al., 2013).

3. Data and metrics

3.1. Dataset

We use the public dataset of Reinecke and Gajos (2014), which contains 398 webpage screenshoots with aesthetics ratings from around 40,000 users via a web based user study (LabintheWild.org). The webpages come with a large variation in terms of aesthetics rating

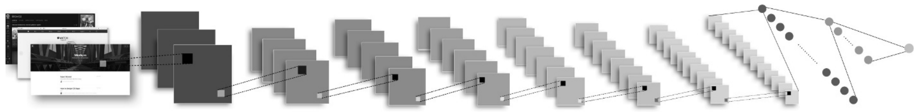


Fig. 1. Architecture of the *Webthetics* deep learning model for webpage aesthetics quantification.

qualities, and were rated with scores between 1 to 9. We randomly divided the dataset into 300 training samples (75.4%) and 98 testing samples (24.6%). The relatively large proportion of testing data could extensively validate the generalization capability of our deep learning model. For pre-processing of the webpage screenshoots, we performed four times downsample from the resolution of 1024×768 to 256×192, and rescaled the intensities into the range of [0,1].

3.2. Results evaluation measurement

We aim to predict webpage aesthetics ratings that have positive linear relationship with the real user ratings. In this case, we employ the Pearson product-moment correlation coefficient (r) as quantitative measurement of the strength and direction of the linear relationship between our results and ground truths. The Pearson coefficient can range from -1 to 1, with -1 indicating perfect negative linear relationship and 1 for perfect positive linear relationship. We also calculated the 95% confidence intervals of the Pearson coefficients with the standard statistical method (Bonett and Wright, 2000). A p-value of lower than 0.05 is considered to indicate statistical significance.

4. Methods

Automatic webpage aesthetics evaluation is a challenging task, given that visual appealing is an ambiguous concept to rigorously define. Even for human users, it still seems quite difficult to explain why someone perceives a webpage design attractive or not. We hypothesize that webpage aesthetics can be related to many aspects of characteristics, for examples, the low-level image statistics (e.g., color and texture), layout (e.g., symmetry and equilibrium), complexity (e.g., number of decomposition regions), and etc. With these considerations, we realize that automatic webpage aesthetics rating is a quite complicated visual computational task in order to produce reliable predictions.

Recent deep learning technologies, especially the CNNs, have achieved broad successes on many extremely challenging image processing tasks (He et al., 2016; Krizhevsky et al., 2012). With an exceeding capability to extract highly representative features from raw pixel intensities, the CNNs are naturally suitable for tasks that are difficult to hand-craft features, such as the problem of webpage aesthetics rating. In these regards, we propose to use deep CNNs to develop our *Webthetics* method.

4.1. Deep learning formulation

In our employed dataset, users have rated the webpages a Likert scale of 1 to 9, i.e., $1 \sim 9$ with 1 standing for the least aesthetically pleasing and 9 standing for the most pleasing. With this aesthetics rating mechanism, we notice that the distance between different rating scores should not be regarded equally. Calculation of the loss between the model prediction and ground-truth should carefully consider this situation. The underlying reason is that differences between rating scores would correlate with different scales of perceived aesthetics. For example, aesthetics ratings of 4 and 5 are "nearer" than ratings of 4 and 7, because scores of 4 or 5 would implicitly fall into the same aesthetics appealing level, whereas a rating of 7 would be considered far more appealing from users perceptive. This observation is crucial for the objective formulation when employing deep learning as a solution. If we employ the traditional cross-entropy loss which formulates the task as a classification problem, the miss-classification predictions into different categories would generate an equal loss. For a concrete example, given a webpage with aesthetics rating of 4, the cross-entropy losses for predicting score of 5 and 7 are the same. However, it would be more reasonable to give them different losses. More specifically, predicting into 7 should result in higher loss compared with predicting into 5, because a prediction rating of 7 over-estimates the aesthetics more heavily.

In this regard, we propose to formulate the webpage aesthetics rating task as a real-valued regression problem other than a classification problem, even though the ratings come with a discrete format. Under this consideration, our target is to predict continuous webpage aesthetics rating scores other than discrete category labels. Within the dataset, each webpage has received aesthetics ratings from many users and each user has rated a set of webpages. In this paper, we employ webpage-user rating pairs to train the CNN which is in essence a supervised machine learning approach. For the testing procedure, the ground truth of each webpage is the average value of all user ratings it has received.

Denoting a webpage screenshoot by x_n, we aim to construct a regression mapping $f: R^2 \to R$ which estimates the aesthetics rating prediction as $\hat{y}_n = f(x_n)$. Given the webpage-user rating pairs (x_n, y_n), $n = 1, \ldots, N$, we formulate the webpage aesthetics rating regression loss as $\ell = \frac{1}{2N}\sum_{n=1}^{N} \|\hat{y}_n - y_n\|_2^2$, which computes the square errors between the rating predictions $\hat{y}_n$ and real rating values y_n from users. Our method is to exploit a deep convolutional neural network to automatically learn the mapping function f, as shown in Fig. 1. Denoting the parameters of the network by W, the overall optimization objective is to minimize the following loss function:

$$\mathcal{L} = \frac{1}{2N}\sum_{n=1}^{N} \|\hat{y}_n - y_n\|_2^2 + \lambda\, \|W\|_2^2 \tag{1}$$

where the second term is the L2-Norm regularization that drives the weights closer to the origin (Goodfellow et al., 2016), and λ is the trade-off parameter. By learning from the webpages and users aesthetics rating data, the network gains the capability to extract representative features directly from the webpage screenshoots. Taking advantage of these high-level features, the deep learning model is able to predict reliable webpage aesthetics ratings.

4.2. Knowledge transfer for effective aesthetics rating

The webpage dataset is relatively small which would bring the risk of over-fitting for training the model. We exploit transfer learning which can set up the model from a good starting point with initialization from a pre-trained network. This strategy effectively enables us to train a deep network with limited data. Intuitively, transfer learning stands by the observation of the generality versus specificity of the neurons in the network. To be specific, the earlier layers tend to contain more generic features (e.g. edge filters and color blobs in the first layer) that are applicable to many image processing tasks; whereas the upper layers progressively become more specific to the particular task.

In transfer learning, the practical approach is that we first have a base network that has been trained on a base dataset towards a base task. When given a new target task, we initialize the layers of the target network with those of the base network, and then fine-tune the whole target network on the target dataset. Existing studies have documented that the transferability of learned knowledge will increase as the distance between the target task and the base task decreases (Yosinski et al., 2014). Inspired by this finding, we are

Table 1
Layers of the deep convolutional neural network.

Layer	Kernel	Stride	Channel
conv1	11 × 11	4	96
pool1	3 × 3	2	96
conv2	5 × 5	1	256
pool2	3 × 3	2	256
conv3	3 × 3	1	384
conv4	3 × 3	1	384
conv5	3 × 3	1	256
pool5	3 × 3	2	256
fc6	–	–	1024
fc7	–	–	512
Regression	–	–	1

interested to repurpose those base models whose tasks share common ground with our target aesthetics evaluation task.

In this regard, we choose the model pretrained on the Flickr Style dataset (containing 80 K images) for image style recognition task as our base network (Karayev et al., 2013). This base model aims at the artistic aspect recognition of photographs, which is more related to our aesthetics task, when compared with other strict object recognition task (Krizhevsky et al., 2012). Furthermore, the network architecture, which is referred as CaffeNet, is quite clear and can be generally adapted. We have transformed the class space of the base models by replacing the last layer (containing 1000 neurons for ImageNet dataset) with a single-neuron layer. In addition, the classification cross-entropy loss is changed into regression loss as we formulated before, by modifying the Softmax loss into the Euclidean loss. We also have reduced the number of neurons in the last two fully-connected layers, to reduce the number of parameters and therefore to further alleviate the risk of over-fitting. To the end, the structure of our *Webthetics* model is listed in Table 1, and the network contains 5 convolutional (conv) layers, 2 max-pooling (pool) layers, 2 fully-connected (fc) layers and a regression layer. Since we are estimating the aesthetics of the image which is a more global or high-level concept, we chose to employ large kernel size for the first convolution layer rather than smaller ones that are often used in image classification tasks.

Interestingly, our base style recognition network is fine-tuned from another base network which has been pretrained on the ImageNet dataset for object recognition task. The ImageNet dataset is quite large scale, containing over 1.2 million natural images from 1000 classes. To explore the transferability of the learned aesthetics related knowledge by the style recognition model, we have also conducted experiments to directly transfer knowlege from the ImageNet base model. Our detailed results regarding this exploration are presented and analyzed in the Experiments Section.

4.3. Learning process

We trained the deep convolutional networks using the standard back-propagation algorithm. The learning rate was initially set as 0.001 and annealed over the training process by dividing a factor of 10 every 2 K iterations. Each iteration employed a batch size of 64; the momentum was set as 0.9; the weight decay (λ) was 0.001. We utilized the dropout strategy (rate = 0.5) in the fc6 and fc7 layers to improve the model's generalization capability.

5. Experiments

5.1. Webpage aesthetics rating prediction

We deployed the trained deep learning model for webpage aesthetic quantification on the testing dataset. The prediction results are shown in Fig. 2. We can observe that the estimated ratings are highly correlated with the user ratings, with a Pearson value of $r = 0.85$, $p < .001$. This reflects that the visual appearance and design factors that affect webpage aesthetics can be successfully represented by the features learned from the CNN model. The deep network has effectively gained knowledge and is capable of extracting highly discriminative representations from the webpage visual information, even though the potential aesthetics factors seems to be subjective and difficult to explicitly define.

We have also analyzed the statistical distribution of the ratings from users and those from our deep learning model, see Fig. 3. This analysis is based on the testing dataset which was randomly sampled from the whole webpage dataset. We show the histograms of the rating scales embedded in different bins. Intuitively, in Fig. 3 left which is the rating distribution of users, we observe that the majority (62%) of the webpages received a score of middle level (i.e.,4 ~ 6 points). Only a small number of webpages obtained an extremely low or high score. Our aesthetics rating predictions also present a similar distribution, see Fig. 3 middle, where 63% of webpages have received a score within the range of 4 ~ 6 points. We statistically fit the histograms into smoothed distributions and overlay them together in Fig. 3 right. It is observed that the statistical distribution of CNN predictions is highly consistent with the users' actual ratings with over 90% overlapping.

In Fig. 4, we illustrate several specific examples of webpage with both the user rating scores and also the aesthetic scores predicted by our Webthetics model. We can find that the ratings predicted by our model highly agree with the users actual ratings. To take a closer look at the examples, the webpages (a) and (b) receive higher aesthetic rating scores and their design seems to be more visual, stylish, and modern with clearer visual hierarchy. In contrast, the webpages (e) and (f) have much lower aesthetic rating scores from both the users and our

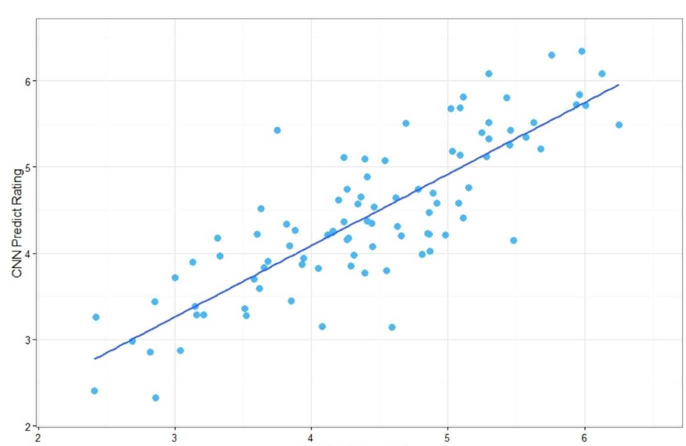


Fig. 2. Webpage aesthetics rating predictions with our deep learning model ($r = 0.85$, $p < .001$).

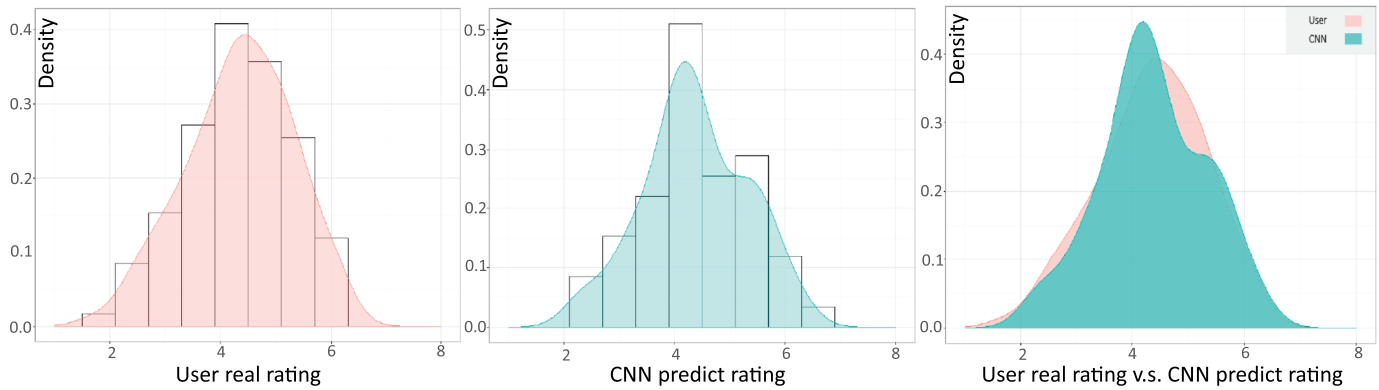
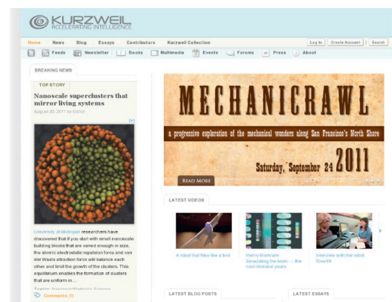


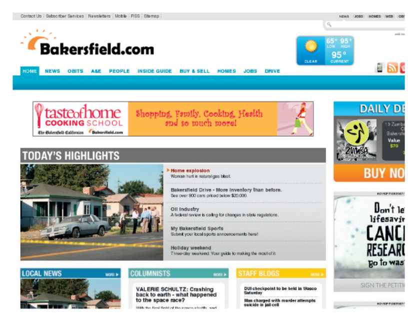
Fig. 3. Analysis of distribution of webpage aesthetics rating predictions from deep learning model.

(a) User average rating 6.13 v.s. CNN prediction 6.09.

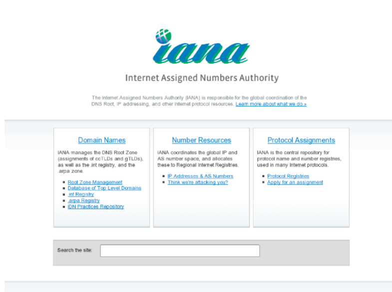


(b) User average rating 5.45 v.s. CNN prediction 5.26.

(c) User average rating 4.27 v.s. CNN prediction 4.18.



(d) User average rating 4.16 v.s. CNN prediction 4.25.



(e) User average rating 3.62 v.s. CNN prediction 3.59.

(f) User average rating 2.69 v.s. CNN prediction 2.98.

Fig. 4. Concrete examples of webpage aesthetics rating predictions from deep learning model.

Fig. 5. Two-dimensional t-SNE embedding visualizations of the webpages of the testing set rated with different scores. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Table 2
Evaluation of Pearson correlations of results using different knowledge transfer states. The values in the parentheses are 95% confidence intervals.

Knowledge	Style recognition	Object recognition
No transfer	0.56 [0.41,0.68]	0.56 [0.41,0.68]
Soft transfer	0.85 [0.79,0.90]	0.83 [0.76,0.88]
Hard transfer	0.83 [0.76,0.88]	0.82 [0.74,0.88]

Table 3
Hand-crafted features for webpage aesthetics quantification with linear regression model.

Factor	Df	Sum Sq	Mean Sq	F value	Pr(> F)
colorfulness	1	13.99	13.99	16.54	5.96e-05
complexity	1	31.07	31.07	36.73	3.69e-09
colorfulness^e	1	12.55	12.55	14.83	1.4e-4
complexity$^{0.4}$	1	14.74	14.74	17.42	3.82e-05
contrast	1	3.14	3.14	3.71	0.055
Residuals	332	280.835	0.8459	–	–

model, and their design seems to be quite textual, complex, cluttered, or even dull and old fashion looking.

We embedded the features in the fc7 layer into a 2D plane for visualization using the t-SNE technique (Maaten and Hinton, 2008), as shown in Fig. 5. The learned high-level features via the deep learning model are highly representative for aesthetics ratings. We divide the testing webpages into four groups (indicated by different colors) according to their aesthetics scores given by users. It is observed that the features of webpages with similar aesthetics rating tend to cluster together. The high-rating webpage features (see red dots) are distinct from those low-rating webpages (see blue dots) in the feature space. This visualization pattern validates that the features are highly correlated with the webpage aesthetics, and hence demonstrates their strong representation capability. In fact, the t-SNE embedding visualizations show the trend that the webpages with higher aesthetics rating are more graphical, visual, and stylish, whereas the ones with lower aesthetics ratings are more textual, complex, or old-fashioned looking.

5.2. Importance of aesthetics knowledge transfer

To analyze the effectiveness, importance and transferability of image style recognition knowledge, we examined the model performance at three different knowledge transfer scales, i.e., *no transfer, soft transfer* and *hard transfer*. More specifically, for the *no transfer* situation, all the weights within the network were initialized randomly from Gaussian distribution $\mathcal{N}(0, 0.01)$. In this setting, the model learned webpage aesthetics rating from scratch. For *soft transfer* network, we initialized the first two convolutional layers from the image style recognition network. For the *hard transfer* network, we transferred all the convolutional layers from the image style recognition network.

The results are shown in Table 2, under the column of style recognition. We list the results of networks with different knowledge transfer scales from the base model pretrained on image style recognition task. It is observed that the *no transfer* model produced significantly far inferior result than the transferred ones ($p < 0.05$). This presents the effectiveness as well as crucial importance of transfer learning in this task, given the limited size of dataset we had for training. The *soft transfer* model and the *hard transfer* model have achieved comparable results. To demonstrate this, we trained another set of networks (i.e., *no transfer, soft transfer, hard transfer*) using the base model which was pretrained on the ImageNet dataset for object recognition task. This was exactly the same model that the aforementioned image style recognition model fined-tuned from. In this way, the two sets of models came with identical network architectures, and we also used the same learning hyper-parameters. The only difference was their initialization base models. In the object recognition column of Table 2, we present the results of networks transferring knowledge from the ImageNet base model. Comparing the results of transferring knowledge from style recognition and object recognition tasks, we find that the results of image style transferred model is just marginally higher than that of the ImageNet transferred model, without observing significant difference. This may be related to the fact that the image style model was also transferred from the ImageNet base model. Overall, the transfer learning technique can help improve the performance of the aesthetics quantification.

5.3. Comparison with hand-crafted features

The original dataset has also provided quantification of colorfulness and visual complexity of the webpages, which were regarded as two most important factors for predicting visual appealing (Reinecke et al., 2013). The colorfulness feature was obtained with a perceptually-based HSV model by comprising a color's hue, saturation and value. The visual complexity feature was determined with space-based decomposition, symmetry, balance and equilibrium. Reinecke et al. (2013) has extensively verified the effectiveness of the colorfulness and complexity features. Following their work, we construct a linear regression model based on these hand-crafted features as the baseline in our experiments. The set up of training and testing data were the same for both the baseline and our deep learning method.

The results of statistical analysis regarding the hand-crafted features for linear regression are listed in Table 3. Besides the colorfulness and complexity, the statistic histogram transformation analysis of these factor's indicates that colorfulness^e and complexity$^{0.4}$ also have

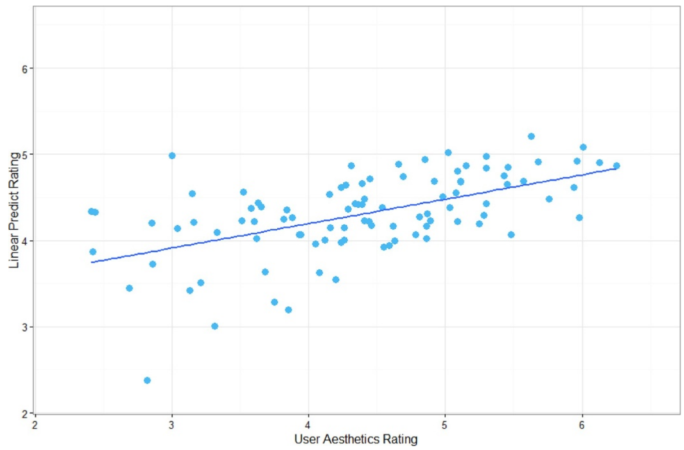


Fig. 6. Aesthetic rating predictions with hand-crafted feature regression model ($r = 0.59$, $p < .001$).

significant effects on the webpage aesthetics rating. In addition, from the perspective of professional web visual design, we understand that the contrast of colorfulness and complexity can influence the webpage aesthetics appealing. To reflect this finding in hand-crafted features and therefore enhance the baseline linear regression model, we introduce the factor of contrast, which was calculated as $contrast = |colorfulness^3 - complexity^3|$ and included into the regression model.

Results of the hand-crafted feature regression model are shown in Fig. 6, where the correlation is $r = 0.59$, $p < .001$. In comparison with the deep learning results in Fig. 2, we find that the rating predictions from CNN are much closer to the actual user aesthetics ratings than those of the hand-crafted feature regression. These results show that the deep learning model has outperformed the baseline method by a significant margin, confirming the effectiveness of deep learning on the webpage aesthetics quantification task. This superior performance can be attributed to the high-level representations that are directly extracted from raw webpages via the deep neural network trained with data.

Moreover, we observe a distinct advantage of deep learning model when evaluating some relatively high or low rated webpages, see examples in Fig. 7. For Fig. 7 (a), the color is simplex but the webpage's layout is well organized. Some users, especially young people, may show preference on these designs and think the contents are easy to understand. The hand-crafted feature regression model produces a score of 5.07 whereas our deep learning model predicts a higher score of 5.71, which is much closer to the user's average rating. On the other hand, for the webpapge in Fig. 7 (b), its colorfulness and visual complexity are strong, however, users did not perceive appealing aesthetics from this design. In this case, the deep learning model gives a low rating of 2.40, whereas the regression model still rates as high as 4.34. From these concrete examples, we can observe that it's difficult for the hand-craft feature approach, although based on visual design rules, to sufficiently describe many other high-level important design factors. In contrast, the deep learning model has the capability to meet this challenge, because the neural network can automatically extract high-level features from raw webpage input based on the training from the user rating data.

5.4. Model's sensitivity to image manipulations

In this section, we empirically study the model's sensitivity to image manipulations or distortions. The idea is that if the responses from the neural network are sensitive to variations regarding to a kind of manipulation, this can imply that our deep learning model has gained some insights for it. The setting of this empirical experiments are inspired by some of techniques our designers adapted from the well-established visual design principles (Arnheim, 1974; Dondis, 1974; Kadavy, 2011) and often used in their working process to analyze the aesthetics of a screen or a webpage. The manipulation factors came from the review of past literature (Lavie and Tractinsky, 2004; Tuch et al., 2010) and particularly inspired by Bauerly and Liu (2006, 2008) who developed a computational model for webpage aesthetic judgement. More specifically, they designed experiments using artificially generated images with different numbers of black and white geometric shapes varying in size and compositional layouts, in order to systematically manipulating different design attributes, such as symmetry, balance or number of the design elements. They demonstrated that computational results from the abstract black and while geometric images were consistent with experiments that used actual webpages. Therefore, we also tried to artificially manipulate the webpage design with geometric shapes.

(a) Users' rating 6.01, hand-crafted feature regression model rating 5.08, deep learning model rating 5.71.

(b) Users' rating 2.41, hand-crafted feature regression model rating 4.34, deep learning model rating 2.40.

Fig. 7. Comparison of webpage aesthetics ratings between deep learning model and hand-crafted feature regression.

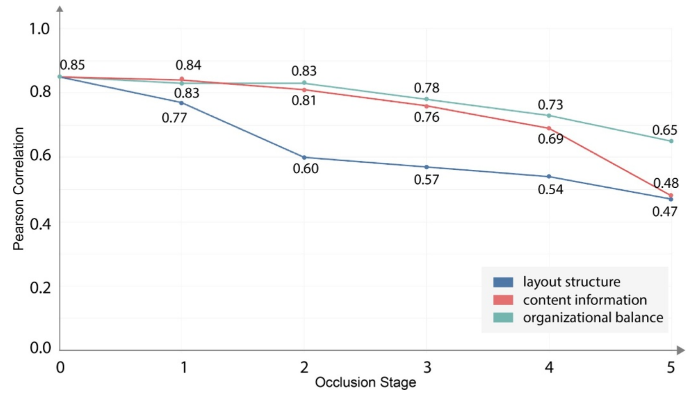


Fig. 8. Empirical study of deep learning model's sensitivity to important web manipulations. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Fig. 9. Example of the occluded webpage in setting-1, from top left to bottom right, are the original webpage and its 100 10 × 10 to 50 × 50 random block occlusions, respectively. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

In the first setting, we included 100 small block occlusions with size of from 10 × 10 to 50 × 50 pixels. The occlusion locations were randomly selected in the webpages, see Fig. 9. The prediction results are presented via the blue line of Fig. 8. We find that the correlation with user ratings decreases rapidly when we add the distortions that impact the page composition elements. This observation indicates that the learned deep learning model is very sensitive to the component organizations as a whole, which is a quite high-level concept of web aesthetics and very difficult to explicitly define.

In the second setting, we occluded the webpages using single black blocks with from 100 × 100 to 500 × 500 pixels in size. The block had an uniform probability to occlude any content within the webpage, such as the text, image, background, and etc., see Fig. 10. The prediction results are presented via the red line of Fig. 8. We can observe that

Fig. 10. Example of the occluded webpage in setting-2, from top left to bottom right, are the original webpage and its 100 × 100 to 500 × 500 random block occlusions, respectively.

Fig. 11. Example of the occluded webpage in setting-3, from top left to bottom right, are the original webpage and its 10% to 50% occlusions, respectively.

Fig. 12. Example of webpages strengthening high and low spatial frequency contents.

when we occlude relatively small regions (i.e.,100×100 and 200×200), the aesthetics rating predictions remain stable. As we occlude more and more contents within the webpages, the correlation with user ratings gradually decreases. This observation demonstrates that the deep learning model can be sensitive to the content contrast included in the webpage.

In the third setting, we occluded different percentages, i.e., 10%, 20%, 30%, 40%, 50%, of the webpages from left to right, see Fig. 11. Our underlying intention is to influence the webpage's organizational balance and observe the network's responses to this variation. The results are shown via the green line of Fig. 8. We can find that the deep learning model is also sensitive to the organizational balance of the webpage to some extent. However, it is not as sensitive as the previous two manipulations.

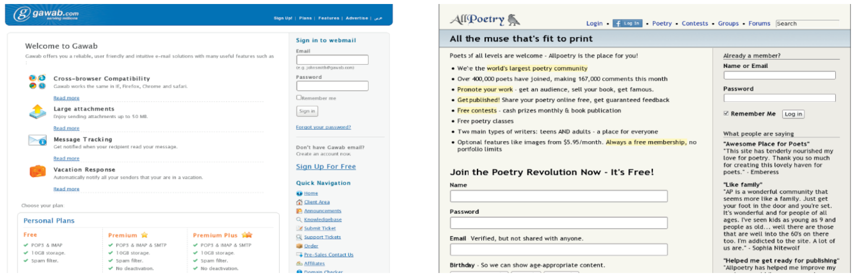
Finally, we explore how the high frequency and low frequency content within the webpage affect the model's response. We respectively employed high-pass and low-pass filters to process the webpages, with examples shown in Fig. 12. The high frequency pass filter enhances details whereas the low frequency pass one blurs the webpage. The high spatial frequency contents mean more fine-grained or local visual patterns, and the low spatial frequency contents mean global visual patterns. For experimental results, the aesthetics rating predictions had a correlation of 0.75 with the user ratings when we enhanced the high frequency content of the webpage, and a correlation of 0.80 for the low spatial frequency enhancement condition.

6. Discussions

Since the World Wide Web was born decades ago, the number of webpages has been growing exponentially, and currently there exists almost 50 billion websites (Douneva et al., 2015), with many of them offering quite similar contents. Given the enormous amount of information on the Internet, people just spend only a few seconds when they first visit a website (Robins and Holmes, 2008). We have reached an era where the functional and usability aspects of a website are almost taken for granted as a basic requirement. Attracting users with beautiful aesthetic design is becoming increasingly important nowadays. With the advancement of analytic technology, there are many tools which have been developed to evaluate webpages from different perspectives, such as W3C HTML Validator, Google Analytics, User Testing.com, etc. Hoffmann and Krauss (2004). Unfortunately, there is still not a prime tool or mature algorithm for evaluating webpages dedicated to the aesthetics of web visual design. Our ultimate goal in this work is to solve this problem and fill this gap.

The architecture of CNN was inspired by the neuroscience research and its connection patterns between different layers of artificial neurons resemble the organization of human visual cortex. Hubel and Wiesel (1962) discovered the existence of different types of neuron in the visual cortex with different size of receptive fields, namely, simple cells with smaller receptive field only react to smaller visual field and simple visual features, such as lines or edges, whereas the complex cells with larger receptive field react to more complex visual patterns in a large visual field. Inspired by human visual system, CNNs architecture also uses restricted receptive field for artificial neurons, and a hierarchy of layers which progressively extract more and more abstract visual features, i.e., hierarchical feature representation. This complex multi-stage/ hierarchical architecture of visual information processing has been the key for the superior performance of image and object recognition task. Our research further demonstrates it can also learn the abstract aesthetic judgement of web page through the learned complex and hierarchical visual representations of the webpages starting from pixels.

Regarding the nature of the dataset where around 40,000 user responses were associated with 398 webpages, we chose to work with the



(a) Users' average rating 3.51 v.s. CNN prediction 4.36.

(b) Users' average rating 2.86 v.s. CNN prediction 4.95.

Fig. 13. Examples of mis-matching cases between the users' ratings and the deep learning model predictions.

average rates of these multiple responses to train the neural network. The reason was that our study aimed to explore a deep learning based tool which learns the general principle of webpage design, and the average rating of users would exactly reflect a relative objective estimation of the aesthetic design quality. Nonetheless, the dataset presented quite complicated variances in relationship with various factors, such as the age, sex, country, education, etc. It would be an interesting future work direction to investigate the correlation of rating variance and characteristics of user groups, which would be advisable for specialized webpage design targeting specific user groups.

Our extensive experimental results have validated that our deep learning technique can be applied effectively for automatic webpage aesthetics evaluation. The predictions from our *Webthetics* model are highly correlated with the actual user ratings, and have outperformed the state-of-the-art hand-crafted feature based model by a large margin, i.e., 0.85 v.s. 0.59. Our proposed approach is also superior to previous deep learning based method of Khani et al. (2016). Coincidently both papers used the same dataset, we applied the evaluation metric used in (Khani et al., 2016) onto our results to obtain the direct comparison between two methods. In this way, our testing error is 20.41%, which is significantly lower than 34.15% reported in that paper. Our end-to-end training practice and regression formulation are the major contributors to the performance improvement.

Furthermore, aesthetic valuation strongly associates with time periods and historical pursuit of web design. The visual culture changes over time and would affect the design of pictorial representations including the online user interfaces, i.e., webpages (Silvennoinen and Jokinen, 2016). If we stick to hand-crafted features for computational webpage aesthetics, researchers have to periodically update the computational elements to match the latest public taste of web aesthetics. This process requires great efforts to define the aesthetic elements and validate them. In contrast, the deep learning model can automatically learn representations from the latest well-designed webpages. It would be much more efficient and effective to update the evaluation tool as the web visual culture evolves.

Another important benefit of CNN is that it does not assume any prior knowledge of a domain other than its above mentioned convolutional architecture. All the features and their associated weights are all learned from the training data through backpropagation and gradient decedent. Because aesthetic perception of a webpage is a complex visual process and far from being clearly understood, any research or models only focus on limited number of visual features would probably too simplistic to account for our complex aesthetic perception. Namely, previous models using handcrafted features, such as complexity or color, would fail to capture all the richness or complexity of the aesthetic perception.

Our deep learning model could help UI or user experience designers during the their everyday design process, by providing an objective aesthetics ratings of the web pages or interfaces that they are working on. In addition, the model can assist designers to adjust and improve their design, by leveraging its sensitiveness to important high-level manipulation factors. We further visualize typical examples of webpages that receive different scales of aesthetics scores from our system in Fig. 5. Based on these visualizations, the designers can directly perceive aesthetics of the webpage, and compare with the high-rating ones, which would greatly benefit improvement of the webpage design.

We analyzed the failure cases that our model predictions mis-matched with users' ratings. We observed that the mis-matchings were related with the figure-text balance and the detailed text information within the webpage. Some human low-rated webpages have few figures in the design. For examples in Fig. 13, users regard the full-text webpage design as not that aesthetics appealing. However, the deep learning model implicitly composes multiple design factors and predicts a higher aesthetics rating. Interestingly, for Fig. 13(b), we conjecture that the shadows on some texts would discourage users to give a high rating. However, the deep learning model would focus on high-level or global features while pay less attention on detailed text contexts. All these interesting observations are towards high-level design concepts and cannot be ideally modeled by traditional aesthetics computation methods such as modeling complexity and colorfulness. Furthermore, these interactions with the deep learning based aesthetics evaluation tool would help us understand how people perceive aesthetics and encourage us to re-think or re-create our design protocol.

Finally, we would like to view our work in this paper from a broader perspective. With recent compelling successes of artificial intelligence (AI), an increasing number of research institutes and companies have started to integrate the emerging AI technologies to the classical visual design area. Novel works, such as thegrid.io, have been proposed to create webpage design automatically using AI approach. Our research in this paper also serves as an important attempt to push forward the AI influence on web design area. Our proposed deep learning model learns to understand aesthetics by seeing hundreds of webpages with visual appealing ratings from users. Ultimately, we believe that the computer is able to develop the capability to evaluate aesthetics of any products just like what human-beings can do.

7. Conclusion and future work

In conclusion, we have presented a novel automatic method to quantitatively evaluate webpage aesthetics. We have demonstrated that our deep learning model trained with knowledge transfer is able to effectively provide aesthetics predictions with high correlations with the real user ratings. The proposed method has significant design implications and it can serve as an efficient means for providing objective aesthetics evaluation during the design process. In addition, our work in this paper contributes to bridge the human computer interaction researches with the recent deep learning evolution.

For future work, we are interested to integrate the user backgrounds or demography information (e.g., age, gender, education, etc.) into our deep learning model. This idea arises from previous studies suggesting that users with different backgrounds might have different aesthetics tastes (Martindale et al., 1990; Reinecke and Gajos, 2014). With this following work, we aim to reveal individual visual preference and then achieve personalized aesthetics rating prediction with the deep learning approach.

Acknowledgment

This work was supported by a grant from the Research Grants Council of HKSAR under General Research Fund (Project no. 14225616).

References

Arnheim, R., 1974. Art and Visual Perception. Univ of California Press.
Bauerly, M., Liu, Y., 2006. Computational modeling and experimental investigation of effects of compositional elements on interface and design aesthetics. Int. J. Hum. Comput. Stud. 64 (8), 670–682.
Bauerly, M., Liu, Y., 2008. Effects of symmetry and number of compositional elements on interface and design aesthetics. Intl. J. Hum. Comput. Interact. 24 (3), 275–287.
Bonett, D.G., Wright, T.A., 2000. Sample size requirements for estimating Pearson, Kendall and Spearman correlations. Psychometrika 65 (1), 23–28.
Datta, R., Joshi, D., Li, J., Wang, J. Z., 2006. Studying aesthetics in photographic images using a computational approach. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 288–301.
Dhar, S., Ordonez, V., Berg, T.L., 2011. High level describable attributes for predicting aesthetics and interestingness. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1657–1664.
Dondis, D.A., 1974. A Primer of Visual Literacy. MIT Press.
Douneva, M., Jaron, R., Thielsch, M.T., 2015. Effects of different website designs on first impressions, aesthetic judgements and memory performance after short presentation. Interact. Comput. iwv033.
Galitz, W.O., 2007. The Essential Guide to User Interface Design: An Introduction to GUI Design Principles and Techniques. John Wiley & Sons.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. http://www.deeplearningbook.org.
Guindon, R., 1990. Designing the design process: exploiting opportunistic thoughts. Hum. Comput. Interact. 5 (2), 305–344.
Harrison, L., Reinecke, K., Chang, R., 2015. Infographic aesthetics: designing for the first impression. Proceedings of the Thirty-Third Annual ACM Conference on Human Factors in Computing Systems. ACM, pp. 1187–1190.
Hassenzahl, M., 2004. Beautiful objects as an extension of the self: a reply. Hum. Comput. Interact. 19 (4), 377–386.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
Heer, J., Bostock, M., 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 203–212.
Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507.
Hoffmann, R., Krauss, K., 2004. A critical evaluation of literature on visual aesthetics for the web. Proceedings of the Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries. South African Institute for Computer Scientists and Information Technologists, pp. 205–209.
Hubel, D.H., Wiesel, T.N., 1962. Receptive fields, binocular interaction and functional architecture in the Cat's visual cortex. J. Physiol. (Lond.) 160 (1), 106–154.
Hubel, D.H., Wiesel, T.N., 1979. Brain Mechanisms of Vision. WH Freeman.
Jaarsveld, S., Leeuwen, C., 2005. Sketches from a design process: Creative cognition inferred from intermediate products. Cognit. Sci. 29 (1), 79–101.
Kadavy, D., 2011. Design for Hackers: Reverse Engineering Beauty. John Wiley & Sons.
Kang, L., Ye, P., Li, Y., Doermann, D., 2014. Convolutional neural networks for no-reference image quality assessment. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1733–1740.
Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., Winnemoeller, H., 2013. Recognizing image style. arXiv:1311.3715.
Khani, M.G., Mazinani, M.R., Fayyaz, M., Hoseini, M., 2016. A novel approach for website aesthetic evaluation based on convolutional neural networks. Proceedings of the Second International Conference on Web Research (ICWR). IEEE, pp. 48–53.
Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C., 2016. Photo aesthetics ranking network with attributes and content adaptation. Proceedings of the European Conference on Computer Vision. Springer, pp. 662–679.
Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 1097–1105.
Lavie, T., Tractinsky, N., 2004. Assessing dimensions of perceived visual aesthetics of web sites. Int. J. Hum. Comput. Stud. 60 (3), 269–298.
Lawson, B., 2006. How Designers Think: The Design Process Demystified. Routledge.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.
Lindgaard, G., Dudek, C., 2003. What is this evasive beast we call user satisfaction? Interact. Comput. 15 (3), 429–452.
Lindgaard, G., Dudek, C., Sen, D., Sumegi, L., Noonan, P., 2011. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. ACM Trans. Comput. Hum. Interact. (TOCHI) 18 (1), 1.
Lindgaard, G., Fernandes, G., Dudek, C., Brown, J., 2006. Attention web designers: you have 50 milliseconds to make a good first impression!. Behav. Inf. Technol. 25 (2), 115–126.
Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.Z., 2014. Rapid: rating pictorial aesthetics using deep learning. Proceedings of the Twenty-Second ACM International Conference on Multimedia. ACM, pp. 457–466.
Lu, Y., Tan, B., Wang, Y., 2013. Web aesthetics: how does it influence the sales performance in online marketplaces. Proceedings of the International Conference on Information Systems, ICIS. Milano, Italy. December 15–18, 2013
Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-sne. J. Mach. Learn. Res. 9 (Nov), 2579–2605.
Marchesotti, L., Perronnin, F., Larlus, D., Csurka, G., 2011. Assessing the aesthetic quality of photographs using generic image descriptors. Proceedings of the International Conference on Computer Vision. IEEE, pp. 1784–1791.
Martindale, C., Moore, K., Borkum, J., 1990. Aesthetic preference: anomalous findings for Berlyne's psychobiological theory. Am. J. Psychol. 103 (1), 53–80.
Michailidou, E., Harper, S., Bechhofer, S., 2008. Visual complexity and aesthetic perception of web pages. Proceedings of the Twenty-Sixth Annual ACM International Conference on Design of Communication. ACM, pp. 215–224.
Nishiyama, M., Okabe, T., Sato, I., Sato, Y., 2011. Aesthetic quality classification of photographs based on color harmony. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 33–40.
Park, S.-e., Choi, D., Kim, J., 2004. Critical factors for the aesthetic fidelity of web pages: empirical studies with professional web designers and users. Interact. Comput. 16 (2), 351–376.
Reinecke, K., Gajos, K. Z., 2014. Quantifying visual preferences around the world. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 11–20.
Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., Gajos, K.Z., 2013. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 2049–2058.
Robins, D., Holmes, J., 2008. Aesthetics and credibility in web site design. Inf. Process. Manag. 44 (1), 386–399.
Schmidt, K.E., Liu, Y., Sridharan, S., 2009. Webpage aesthetics, performance and usability: design variables and their effects. Ergonomics 52 (6), 631–643.
Silvennoinen, J.M., Jokinen, J.P., 2016. Aesthetic appeal and visual usability in four icon design eras. Proceedings of the CHI Conference on Human Factors in Computing Systems. ACM, pp. 4390–4400.
Thorlacius, L., 2007. The role of aesthetics in web design. Nord. Rev. 28 (1), 63–76.
Tractinsky, N., Cokhavi, A., Kirschenbaum, M., Sharfi, T., 2006. Evaluating the consistency of immediate aesthetic perceptions of web pages. Int. J. Hum. Comput. Stud. 64 (11), 1071–1083.
Treisman, A.M., Gelade, G., 1980. A feature-integration theory of attention. Cognit. Psychol. 12 (1), 97–136.
Tuch, A.N., Bargas-Avila, J.A., Opwis, K., 2010. Symmetry and aesthetics in website design: ITSA mans business. Comput. Hum. Behav. 26 (6), 1831–1837.
Tuch, A.N., Roth, S.P., HornbæK, K., Opwis, K., Bargas-Avila, J.A., 2012. Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in hci. Comput. Hum. Behav. 28 (5), 1596–1607.
Wu, O., Chen, Y., Li, B., Hu, W., 2011. Evaluating the visual quality of web pages using a computational aesthetic approach. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, pp. 337–346.
Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? Proceedings of the Advances in Neural Information Processing Systems. pp. 3320–3328.
Zhang, K., Kong, J., 2010. Exploring semantic roles of web interface components. Proceedings of the International Conference on Machine and Web Intelligence (ICMWI). IEEE, pp. 8–14.
Zheng, X.S., Chakraborty, I., Lin, J.J.-W., Rauschenberger, R., 2009. Correlating low-level image statistics with users-rapid aesthetic and affective judgments of web pages. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 1–10.